APPENDIX I

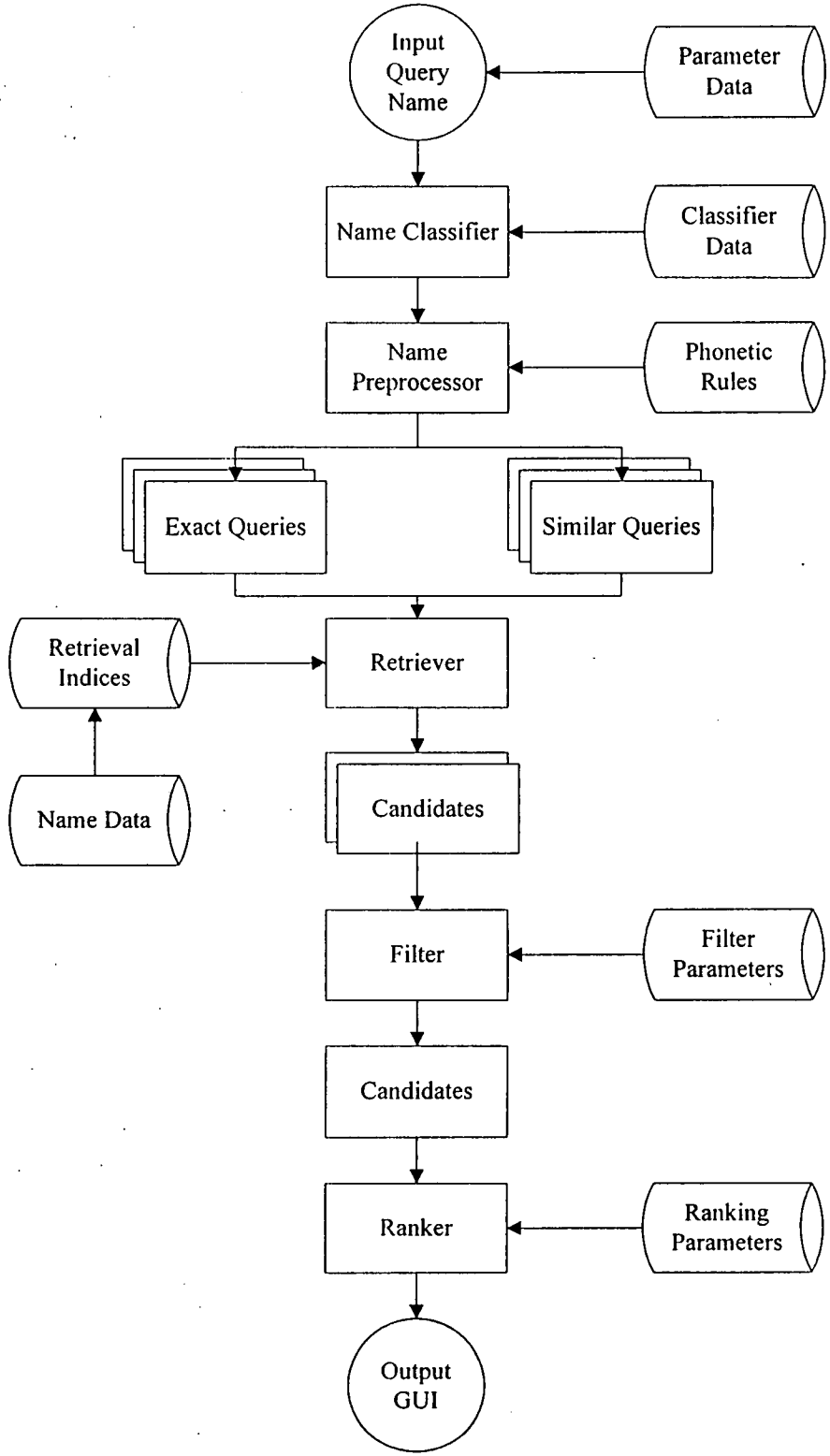
Name Search
Technology Demonstration System
Acceptance Test

February 11, 1998

Language Analysis Systems, Inc.
2214 Rock Hill Road—Herndon, VA—20170

Name Search
Technology Demonstration System
(TDS)
Conceptual Design

11 February 1998



Technology Demonstration System (TDS)

- A phonological name search system
- Technology based on the research results of an 18-month project
- A prototype developed to demonstrate the capabilities of this technology

Characteristics of TDS

- Each query drives the list of returns
  - no *a priori* set or group of names
  - "on-the-fly", automatic process of associating query name with database names
  - allows *multiple* relationships among names
- Pronunciation and spelling contribute to measures of similarity
- System returns ranked list of similar names

Characteristics of TDS (cont.)

- Culture-based rule sets
  - Names are automatically classified as Arabic, Mandarin Chinese, Hispanic or "Other"
  - Arabic, Mandarin Chinese and Hispanic rule sets process names based on automatic classification
    - E.g., Arabic rules: *Qaddafi* ~ *Khaddafy*
  - *All* names are processed by Anglo rules

Key Characteristics of the TDS

- *fast*
- *principled*
- *fully automatic*
- *sorted returns*
- *multicultural*

Scope of the TDS

- Not a full retrieval system
  - retrieves single names (e.g., *Smith*)
- Other factors *not* covered by TDS:
  - Stems and affixes: *Vega ~ Delavega*
  - Dialect: Chinese *Ng = Wu = Huang*
  - Typographical errors: *Jpnes*
  - Perceptual issues: *Polk* misheard as *Holtz*

Technical Overview of the TDS

- Written in Microsoft Visual C++ for Windows NT
- Six-month development period
- Effort concentrated on indexing strategies, search algorithms and ranker
- Minimal effort spent on User Interface
- Currently running on an IBM ThinkPad 770 with a 233 MHz Pentium MMX, 160 Mb of RAM

TDS Acceptance Test

Purpose of today's Acceptance Test:

To determime whether the TDS satisfies the requirements of the *Statement of Work*

Requirements of the TDS

1. The TDS will incorporate a search component using phonetic name search algorithms (IPA exact match and phonetically "similar-to")
2. The TDS will incorporate a name classifier component automatically identifying a query name as a member of a specific culture for which culture-specific name processing rules can be applied
3. The cultures to be implemented in the Name Classifier component are Arabic, Chinese, Hispanic and Other (including Anglo)
4. The TDS will incorporate a rank-ordering component that ranks the results from the search component, a name database and its supporting database management system, and a graphical user interface

Requirements of the TDS (cont.)

5. The individual software components that are directly related to name searching should be designed and written to be modular
6. The TDS will work on a name database consisting of at least three million names
7. The TDS will allow a user to input as a query a name in Romanized form
   - Accepts entry of a single name segment only
   - Accepts input length 2 - 30 characters as valid entry
   - Accepts lower and upper case letters as valid entry
   - Accepts alphabetic characters and apostrophe as valid entry

Requirements of the TDS (cont.)

8. Names that have been classified by the Name Classifier as Arabic, Chinese or Hispanic will be processed by their respective language-specific components, as well as by the English (Anglo)-language components. All names will be processed by the Anglo components.
9. The TDS will rank and display retrieved names in order of phonological similarity to the query name, with Exact Matches displayed at the top of an ordered list
10. The TDS will include a batch processing component
11. The ranking algorithm should be general enough to apply to the full set of names retrieved regardless of whether the technology responsible for the retrieval was IPA exact match or phonetically similar-to

Requirements of the TDS (cont.)

12. The TDS shall begin to display the results of each query against a name database of about three million names within twelve (12) seconds
13. The user shall be able to select for each query whether "similar-to" logic is to be used in the retrieval process
14. The user shall be able to select for each query whether the user will bypass the name classifier and manually specify the culture of the query name
15. The user shall be able to select for each query whether the name classifier is to be used; if not, the query shall be processed as an English (Anglo) name (in addition to the manually specified culture, if any)

Requirements of the TDS (cont.)

16. The user shall be able to select for each query the maximum number of hits to be displayed
17. The reason a name was retrieved (that is, whether the retrieval of a name was due to IPA exact match or phonetically "similar-to" technology) shall be displayed along with the name
18. The TDS display will include the list of hits returned along with each name's six character group number
19. Each option selected shall become the default option and shall apply to all queries until the user changes it
20. The contractor may include other options that satisfy the needs of its developers and other personnel, as long as the default for those options is off

Requirements of the TDS (cont.)

21. The TDS software shall be written with English comments embedded in the code that implements each algorithm
22. The comments shall tie blocks of code (a block of code is one or more sequential lines of code) in each algorithm's implementation to the step being implemented in the deliverable English-language narrative that describes the algorithm
23. The TDS should be designed so that new technology can be easily incorporated as it becomes available
24. The TDS will include a batch processing option for results retrieval

Requirements of the TDS (cont.)

25. The TDS will accept the following input when performing the pre-processing of the client's names database:
    - ASCII text file
    - Allowable characters include letters of the alphabet and the apostrophe (').
      - Lower case letters will be converted to upper case
    - Each record will contain two fields separated by a blank as specified below:
      - Columns 1 through 6 will contain the Group Number; the first character may be either a 'Z' or a number; the remaining 5 characters must contain a digit
      - Column 7 must be blank
      - Columns 8 through 37 will contain the name. The name must be at least 2 characters long, and no longer than 30 characters
    - Client database may exist on 3.5" diskette or CD-ROM
    - Duplicate names will be rejected

Requirements of the TDS (cont.)

26. Each query name and all the names retrieved by the TDS (not to exceed the limit in use at that time) will be saved in an internal file until the user selects to delete it
27. The TDS will also provide the capability to write the saved data onto a hit file resident on diskette(s) when directed by the user